A quasi-orthogonal, invertible, and perceptually relevant time-frequency transform for audio coding

Olivier Derrien, Thibaud Necciari, Peter Balazs

A QUASI-ORTHOGONAL, INVERTIBLE, AND PERCEPTUALLY RELEVANT TIME-FREQUENCY TRANSFORM FOR AUDIO CODING

Olivier Derrien[*]*, Thibaud Necciari*[†] *and Peter Balazs*[†]

[*] Université de Toulon and Laboratoire de Mécanique et d'Acoustique - CNRS, Marseille, France
[†] Acoustics Research Institute - ÖAW, Vienna, Austria
derrien@lma.cnrs-mrs.fr, thibaud@kfs.oeaw.ac.at, peter.balazs@oeaw.ac.at

ABSTRACT

We describe ERB-MDCT, an invertible real-valued time-frequency transform based on MDCT, which is widely used in audio coding (e.g. MP3 and AAC). ERB-MDCT was designed similarly to ERBLet, a recent invertible transform with a resolution evolving across frequency to match the perceptual ERB frequency scale, while the frequency scale in most invertible transforms (e.g. MDCT) is uniform. ERB-MDCT has mostly the same frequency scale as ERBLet, but the main improvement is that atoms are quasi-orthogonal, i.e. its redundancy is close to 1. Furthermore, the energy is more sparse in the time-frequency plane. Thus, it is more suitable for audio coding than ERBLet.

Index Terms— Non-stationary time-frequency transforms, ERB filters, MDCT, Audio coding.

1. INTRODUCTION

State-of-the-art lossy audio codecs use real-valued time-frequency (TF) transforms, typically Modified Discrete Cosine Transform (MDCT) for MP3 and AAC [1]. The motivation is that modeling the auditory perception is more efficient in the TF domain. MDCT is perfectly invertible and has a redundancy 1. In other words, the number of transform coefficients equals the number of samples in the signal. Usually, TF transforms for audio coding are orthogonal bases i.e. the vectors (called atoms here) that define analysis and synthesis operators are orthogonal and span the signal space. These properties are usually associated to a fixed frequency resolution that is not in line with auditory perception (see Sec. 2). Practically, a masking threshold is computed on the uniform frequency grid by interpolating masking thresholds computed in another perceptual frequency scale, which is not optimal.

Perceptual TF transforms have already been proposed (e.g. Gammatone [2]) but they do not achieve perfect reconstruction and generate some redundancy, and thus are not suitable for audio coding. In [3], it was proposed to perform a decomposition of the signal on a union of MDCTs, which is well suited for audio coding but the frequency scale is not perceptually-motivated. Recently, the ERBLet transform was proposed [4]. Its frequency resolution is matched to the Equivalent Rectangular Bandwidth (ERB) scale and it achieves perfect reconstruction as long as the redundancy is larger than 1. In this paper, we propose a real-valued variant of the ERBLet called ERB-MDCT, more suitable for audio coding. The frequency scale still follows the ERB scale, but the analysis and synthesis sets of atoms are nearly orthogonal bases, which means that the redundancy is close to 1.

This paper is organized as follows: First, we briefly describe the ERB scale and the ERBLet. Then, we describe the ERB-MDCT, and give some implementation details. Finally, we compare it to a standard MDCT and ERBLet in terms of orthogonality, redundancy and TF energy localization. We also provide TF images obtained with a real audio signal.

2. THE ERB SCALE AND THE ERBLET

The peripheral auditory system can be modeled as bank of bandpass filters usually described by their equivalent rectangular bandwidth (ERB). The ERB (in Hz) of the auditory filter centered at frequency f (in Hz) is [5]:

$$\Delta_f(f) = 24.7 + \frac{f}{9.265} \tag{1}$$

The full range of audible frequencies (20 Hz–20 kHz) can be modeled as a juxtaposition of 39 bandpass filters whose center frequencies f_b, $b \in \{1 \cdots 39\}$, are given by [5]:

$$f_b = 228.8455 \left[\exp\left(\frac{b}{9.265} \right) - 1 \right] \tag{2}$$

In [4], a transform with a resolution evolving across frequency has been formulated based on the theory of non-stationary Gabor frames [6]. Specifically, Gaussian windows with bandwidths satisfying equation (1) are constructed in the frequency domain and equidistantly spaced on the ERB scale according to equation (2). The resulting ERBlet transform is computed by applying the set of windows to the Fourier transform of the signal.

This work was supported by the joint French ANR and Austrian FWF project "POTION", refs. ANR-13-IS03-0004-01 and FWF-I-1362-N30, and by the FWF START-project "FLAME", ref. Y-551-N13.

3. DESCRIPTION OF THE ERB-MDCT

3.1. ERB-MDCT basics

The original MDCT has a constant TF resolution [7]. Extensions were proposed, where the TF resolution changes along time [8]. This "time-domain non-stationary" MDCT is actually used in audio codecs like MP3 or AAC (the coder can switch between two resolutions [1]). Basically, ERB-MDCT is a "frequency-domain non-stationary" MDCT that follows the ERB scale. This is achieved by applying a Discrete Cosine Transform (DCT) to a time-domain non-stationary MDCT.

For a given discrete time-domain of length N samples: $n \in \{0 \cdots N-1\}$, any linear TF transform can be defined by two sets of signals $\psi_{p,\tau}[n]$ and $\hat{\psi}_{p,\tau}[n]$ called respectively analysis and synthesis atoms. For a signal x, analysis and synthesis operators can be defined as [9]:

$$x \mapsto X_{p,\tau} = \langle x, \psi_{p,\tau} \rangle \mapsto \hat{x} = \sum_{p,\tau} X_{p,\tau}\ \hat{\psi}_{p,\tau}$$

where $X_{p,\tau}$ are the transform coefficients and $\hat{x}$ is the reconstructed signal. p is a frequency index and τ a time-shift index. For MDCT, we have $\hat{\psi}_{p,\tau} = \psi_{p,\tau}$ and $\hat{x} = x$ (i.e. perfect reconstruction) except on the edges of the time-domain.

In a first step, we focus on the ERB-MDCT synthesis atoms. In a discrete frequency domain: $k \in \{0 \cdots N-1\}$, we define variable-size MDCT atoms, for $p \in \{0 \cdots P\}$:

$$\phi_{p,\tau}[k] = w_p[k]\ \cos\left[\frac{\pi}{N_p}\left(k - k_p + \frac{N_p}{2} + \frac{1}{2}\right)\left(\tau + \frac{1}{2}\right)\right] \tag{3}$$

where N_p is the MDCT size, $\tau \in \{0 \cdots N_p - 1\}$ and the window w_p can be seen as the frequency response of a band-pass filter centered on k_p. The support of w_p is $\{k_p - N_p \cdots k_p + N_p - 1\}$. When k_p and w_p are properly defined, $\{\phi_{p,\tau}\}$ is an orthogonal basis [8]. The ERB-MDCT synthesis atoms are defined as DCT-IV transforms of $\phi_{p,\tau}$:

$$\hat{\psi}_{p,\tau}[n] = \sum_{k=0}^{N-1} \phi_{p,\tau}[k]\ \cos\left[\frac{\pi}{N}\left(k + \frac{1}{2}\right)\left(n + \frac{1}{2}\right)\right] \tag{4}$$

k_p is related to the main oscillation frequency of $\hat{\psi}_{p,\tau}[n]$ which is $\nu_p = \frac{1}{2N}\left(k_p + \frac{1}{2}\right)$. There are N_p atoms in band p, thus the total number of atoms is $N_T = \sum_p N_p$. k_p should be defined such that the frequencies ν_p follow equation (2), and w_p and N_p such that the transform is invertible. However, we can not force the bandwidth of the atoms to follow equation (1) because of the orthogonality constraint.

3.2. Setting the frequency scale

The frequency (in Hz) corresponding to k_p is $f_p = F_s \nu_p = \frac{F_s}{2N}\left(k_p + \frac{1}{2}\right)$, where F_s is the sampling frequency. Ideally, f_p should follow the ERB scale with v bands per ERB (defined by equation (2) with $b = \frac{p}{v}$), with $f_0 = 0$ and $f_P = \frac{F_s}{2} + \frac{1}{2N}$. This is not possible because:

1. For a given value of v, one can usually not find an integer P such that $f_{\frac{P}{v}} = \frac{F_s}{2} + \frac{1}{2N}$ in equation (2).
2. The extreme values $k_p = 0$ and $k_p = N$ correspond respectively to $f_p = \frac{F_s}{4N}$ and $f_p = \frac{F_s}{2} + \frac{F_s}{4N}$.

Thus, we first set P as the closest integer such that $f_{\frac{P}{v}} \approx \frac{F_s}{2} + \frac{1}{2N}$ in equation (2) and then compute k_p using:

$$k_p = N\left(\frac{\exp\left(\frac{p}{9.265\ v}\right) - 1}{\exp\left(\frac{P}{9.265\ v}\right) - 1}\right) \tag{5}$$

which is an approximation of the ERB scale. These real values will be converted to integers later.

3.3. Setting MDCT sizes

The variable-size MDCT is invertible under the conditions:

$$\begin{cases} w_p^2\left[k_p + k\right] + w_p^2\left[k_p + N_p - k\right] = 1 \\ w_p^2\left[k_p - k\right] + w_p^2\left[k_p - N_p + k\right] = 1 \end{cases} \tag{6}$$

for $k \in \{0 \cdots N_p - 1\}$ and

$$w_p\left[k_p - \frac{N_p}{2} + k\right] = w_{p-1}\left[k_{p-1} + \frac{N_{p-1}}{2} - k\right] \tag{7}$$

for $k \in \{-N_p \cdots N_p\}$ [8]. These conditions imply that the second half of w_{p-1} is the "flipped" version of the first half of w_p with respect to a center of symmetry (see figure 1):

$$k_p - \frac{N_p}{2} = k_{p-1} + \frac{N_{p-1}}{2} \tag{8}$$

We know from equation (5) that $k_0 = 0$. Thus, equation (8) leads to:

$$\begin{cases} k_1 &= \frac{1}{2}N_0 + \frac{1}{2}N_1 \\ k_2 &= \frac{1}{2}N_0 + N_1 + \frac{1}{2}N_2 \\ &\vdots \\ k_P &= \frac{1}{2}N_0 + N_1 + \cdots + N_{P-1} + \frac{1}{2}N_P \end{cases} \tag{9}$$

Solving this system for k_p defined by equation (5) should lead to a suitable sequence N_p. However, there might be an infinite set of solutions (because the system is under-determined) or no solutions at all (because only even integer and increasing sequences N_p are acceptable). In Section 3.6, we propose a heuristic to solve this problem.

3.4. Setting MDCT windows

The perfect-reconstruction conditions (6) and (7) are verified for the following window:

$$w_p[k] = \begin{cases} 0 & k \in \left\{k_p - N_p \cdots k_p^{(1)} - 1\right\} \\ \sin\left[\frac{\pi\left(k - k_p^{(1)} - \frac{1}{2}\right)}{2N_{p-1}}\right] & k \in \left\{k_p^{(1)} \cdots k_p^{(2)} - 1\right\} \\ 1 & k \in \left\{k_p^{(2)} \cdots k_p - 1\right\} \\ \cos\left[\frac{\pi\left(k - k_p - \frac{1}{2}\right)}{2N_p}\right] & k \in \{k_p \cdots k_p + N_p - 1\} \end{cases} \tag{10}$$

with:

$$\begin{cases} \delta_p & = & \frac{1}{2}(N_p - N_{p-1}) \\ k_p^{(1)} & = & k_p - N_p + \delta_p = k_p^{(1)} \\ k_p^{(2)} & = & k_p^{(1)} + N_{p-1} \\ k_p & = & k_p^{(2)} + \delta_p \end{cases} \quad (11)$$

Such a window is illustrated on figure 1.

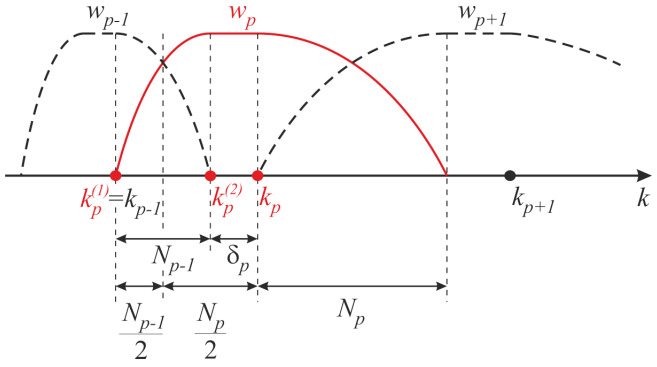

Fig. 1. MDCT window w_p allowing perfect-reconstruction.

3.5. Analysis and synthesis atoms

Equation (4) defines the synthesis atoms. One can see from the system (10) that the union of $w_p[k]$ is defined for $k \in \{-N_0 \cdots N + N_P - 1\}$. Thus, the analysis atoms are:

$$\psi_{p,\tau}[n] = \sum_{k=-N_0}^{N+N_P-1} \phi_{p,\tau}[k] \cos\left[\frac{\pi}{N}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}\right)\right] \quad (12)$$

Perfect reconstruction on variable-size MDCT is achieved when adjacent windows overlap, i.e. for $k \in \{0 \cdots N-1\}$. This justifies the boundaries of the sum in equation (4).

3.6. Implementation details

One can see from equations (3) and (12) that the ERB-MDCT analysis operator is equivalent to:

1. Apply a DCT-IV to the whole signal in the time domain.
2. Apply variable-length MDCT to DCT-IV coefficients.

The synthesis operator follows the reverse scheme. DCT-IV and MDCT can be efficiently implemented using the FFT algorithm. Practically, in an audio codec, the signal should be segmented in overlapping frames that should be processed separately. To minimize the final redundancy, one should use long frames and short overlapping sections.

Computing valid sequences N_p and k_p is not a simple problem (see Section 3.3). We propose a simple heuristic that works for most values of N and v:

1. Compute the (real) target values for k_p as in Section 3.2.
2. Set $N_0 = k_1$.
3. Compute N_p by finding the unique solution to (9).
4. Round each N_p to the nearest even integer.
5. Compute the final integer values of k_p using (9).

The system (9) implies that $2 \leq N_0 \leq k_1$. Thus, no valid solution can be found for $k_1 < 2$. Point 2. comes from the fact that N_p must be an increasing sequence, and we found out that this is always verified when choosing $N_0 = k_1$. However, the final value of k_P may not be equal to N. This can be tackled by iteratively initializing the heuristic with a value of N slightly different from the target value, and stop when k_P matches the target value.

4. DISCUSSION AND EXPERIMENTATION

4.1. Orthogonality and redundancy

We wish that ERB-MDCT analysis and synthesis atoms are each orthogonal sets. For $p = 1 \ldots P-1$, atoms are orthogonal because variable-size MDCT atoms and DCT-IV atoms are orthogonal. However, for $p = 0$ and P, variable-size MDCT atoms are computed from DCT-IV coefficients that are symmetric with respect to 0 and N (see equation (12)). This gives physically-relevant transform coefficients but breaks the orthogonality. Therefore, ERB-MDCT is quasi-orthogonal.

v	K	ERB-MDCT redundancy	ERBLet redundancy
1	43	1.0552	3.6023
2	86	1.0276	7.2922
3	128	1.0186	10.9719
4	171	1.0142	14.6672

Table 1. Number of frequency bands (K) and redundancy as a function of bands per ERB (v). $N = 4096$.

ERB-MDCT redundancy (equal to N_T/N) is always larger than 1 and depends on the discretization of the ERB-scale. On table 1, we give the redundancy as a function of v, for $N = 4096$, for ERB-MDCT and ERBLet (in the painless case i.e. straightforward synthesis). The ERB-MDCT and ERBLet redundancy can not *a priori* be compared, because ERBLet represents positive and negative frequencies with complex coefficients, while ERB-MDCT represents positive frequencies with real coefficients. But in the case of real-valued signals, the comparison is meaningful because of Hermitian symmetry in ERBLet. One can observe that redundancy is close to 1 in ERB-MDCT and much higher in ERBLet. Redundancy decreases with respect to v in ERB-MDCT but increases in ERBLet. This is because ERBLet bandwidths follow equation (1) and do not depend on v. Thus, the overlap between bands increases with v, which is not the case with ERB-MDCT. Practically, audio coding

requires that partials in pitched sounds are resolved, thus a sufficiently high frequency resolution is required (typically $v = 3$), which corresponds to a neglectable redundancy for ERB-MDCT (+2%), whereas ERBLet redundancy is definitely inappropriate for compressive coding.

4.2. Energy localization in time and frequency domains

In this section, we compare the energy localization of synthesis atoms between standard MDCT, ERBLet and ERB-MDCT for $N = 4096$. For ERBLet and ERB-MDCT, we chose $v = 1$, i.e. 43 bands (negative-frequencies in the ERBLet are discarded). We focus on atoms that are approximately centered on $\frac{N}{2}$ and oscillate around 1000 Hz (where the sensitivity of the hearing system is maximal). This corresponds either to $p = 16$ or $p = 17$. For the standard MDCT, we chose the same frequency resolution at 1000 Hz as with ERB-MDCT. This corresponds to 160 bands and either to $p = 7$ or $p = 8$.

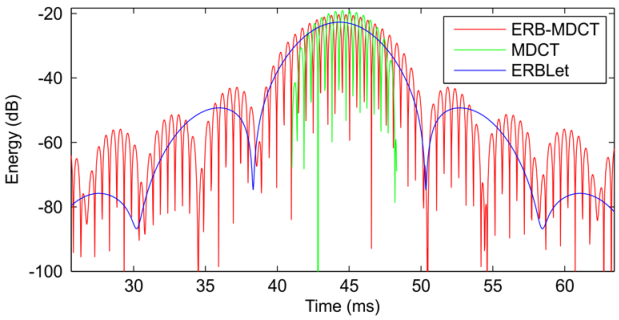

Fig. 2. Energy in the time domain ($N = 4096$, $F_s = 44.1$ kHz) of TF atoms oscillating at 1000 Hz for MDCT (160 bands, $p = 8$), ERBLet and ERB-MDCT (43 bands, $p = 16$).

On figure 2, we plot the energy of atoms in the time domain. One can observe that energy oscillates with cosine-modulated atoms (MDCT, ERB-MDCT), while it is smooth with complex-modulated atoms (ERBLet). MDCT atoms are compactly supported in the time domain (320 samples, i.e. 7.2 ms), whereas others are not. Thus, energy is best localized with MDCT. Furthermore, energy decays faster with ERBLet than with ERB-MDCT: -3 dB at first lobe (time delay: 8.4 ms), and -15 dB at second lobe (time delay: 16.7 ms).

On figure 3, we plot the spectrum of previously-described atoms. As MDCT atoms are compactly-supported in the time domain, their energy decays slowly in the frequency domain. In contrast, ERB-MDCT and especially ERBLet are much more selective. Both follow the ERB-scale but the central frequency is slightly shifted to the right with ERB-MDCT because of the modified ERB scale (see (5)). Furthermore, ERBLet atoms are compactly supported in the frequency domain, which is not the case with others: The attenuation in the stop-band is about -60 dB for ERB-MDCT and -20 dB for MDCT. Thus, energy is better localized in the frequency domain with ERBLet. One can also notice that the bandwidths of ERBLet atoms are broader than those of ERB-MDCT. This is due to the fact that ERBLet atoms are optimized both on ERB center frequencies and bandwidths, whereas ERB-MDCT atoms are optimized only on ERB center frequencies.

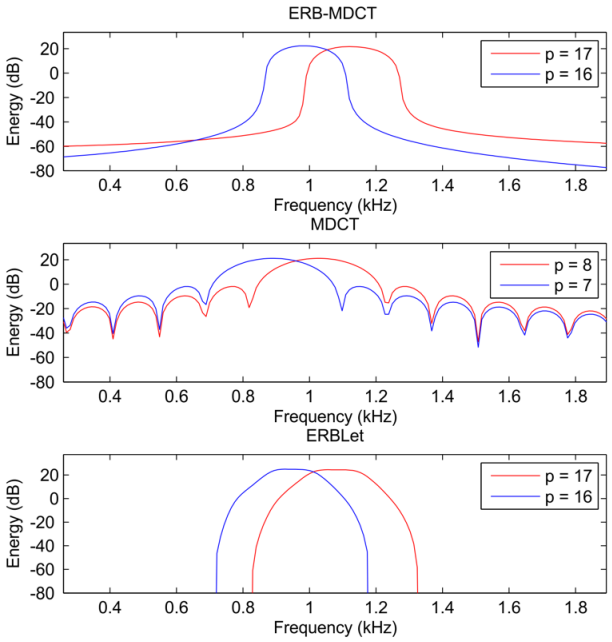

Fig. 3. Energy in the frequency domain ($N = 4096$, $F_s = 44.1$ kHz) of TF atoms oscillating at 1000 Hz, for MDCT (160 bands), ERBLet and ERB-MDCT (43 bands).

4.3. Time-frequency images for a real audio signal

In this section, we compare TF images obtained for a real audio signal: The beginning of "Tom's Dinner" by Suzanne Vega. N equals the length of the audio excerpt and $F_s = 44.1$ kHz. We set $v = 3$ for ERB-MDCT and ERBLet (i.e. 128 bands, keeping only positive frequencies in ERBLet). We also apply a MDCT with the same frequency resolution at 1000 Hz, i.e. with 500 bands. We use the implementation of ERBLet available in the LTFAT 2.0 Toolbox for Matlab (http://ltfat.sourceforge.net/). We also provide online (http://potion.cnrs-mrs.fr/eusipco15.html) an implementation of ERB-MDCT for Matlab. The ERB-MDCT, MDCT and ERBLet TF images are plotted on figure 4.

Between ERB transforms and MDCT, energy spreading is clearly different in the frequency domain: With MDCT, most coefficients in the upper 3/4th represent low-energy information, while high-energy partials are concentrated in the lower 1/4th. With ERB transforms, high frequencies are "compressed" in the upper part, and partials are more salient.

Between ERB-MDCT and ERBLet, the main difference

is that the partials are broader in frequency with ERBLet, because ERBLet bandwidths are wider. Then, partials might be unresolved, especially in high frequencies. In other words, TF representation is more sparse with ERB-MDCT, which is desirable for audio coding: More zero (or zero-quantized) coefficients require less coding bits. Finally, one can observe that the TF image is smoother with ERBLet. This comes from the fact that ERB-MDCT is based on MDCT, which is not shift-invariant in time. This generates local oscillations of energy in the TF plane [10].

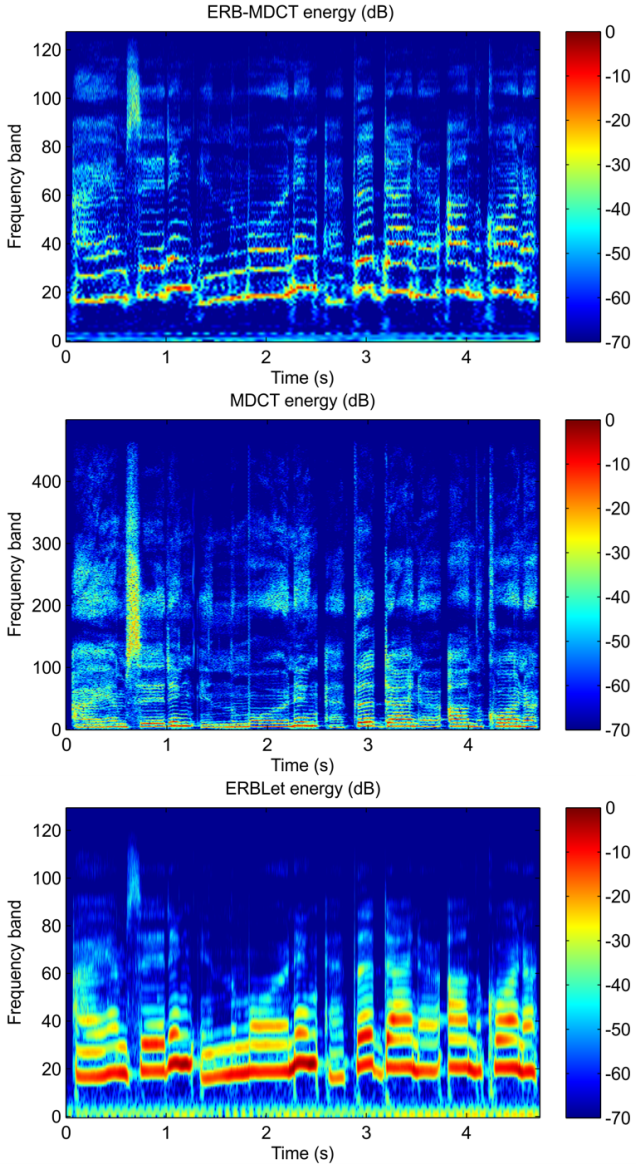

Fig. 4. Time-frequency image of "Tom's Dinner" by Suzanne Vega with ERB-MDCT (128 band), MDCT (500 band) and ERBLet (128 band).

5. CONCLUSION

We proposed a real-valued perfectly-invertible TF transform, inspired by ERBLet, but close to a basis. It was conceived as a trade-off between an efficient modeling of the hearing system and constraints specific to audio coding: Redundancy close to 1, sparse representation in the transform domain, and low computational cost. However, the localization of energy in time and frequency domains is not as good as with ERBLet. In a future work, we will use this transform in a real audio codec and evaluate its performance compared to a MDCT.

REFERENCES

[1] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, Wiley, 2007.

[2] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acust. united Ac.*, vol. 88, no. 3, pp. 433–442, 2002.

[3] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Tr. ASLP*, vol. 16, no. 8, pp. 1361–1372, 2008.

[4] T. Necciari, P. Balazs, N. Holighaus, and P. Søndergaard, "The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 498–502.

[5] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[6] P. Balazs, M. Dörfler, N. Holighaus, F. Jaillet, and G. Velasco, "Theory, implementation and applications of nonstationary Gabor frames," *J. Comput. Appl. Math.*, vol. 236, no. 6, pp. 1481–1496, 2011.

[7] J.P. Princen, A.W. Johnson, and A.B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. ICASSP*, Dallas, TX, April 1987, pp. 2161–2164.

[8] C. Herley, J. Kovacevic, K. Ramchandran, and M. Vetterli, "Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms," *IEEE Tr. SP*, vol. 41, no. 12, pp. 3341–3359, Dec. 1993.

[9] K. Gröchenig, *Foundations of time-frequency analysis. Applied and numerical harmonic analysis*, Birkhäuser, 2001.

[10] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: Exact results and applications to coding artifacts reduction," *IEEE Tr. SAP*, vol. 12, no. 3, pp. 302–312, May 2004.